Distributional Model Equivalence for Risk-Sensitive Reinforcement Learning

Tyler Kastner, Murat Erdogdu, Amir-massoud Farahmand

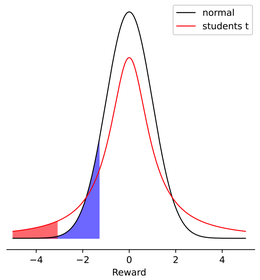
tkastner@cs.toronto.edu, erdogdu@cs.toronto.edu, farahmand@cs.toronto.edu

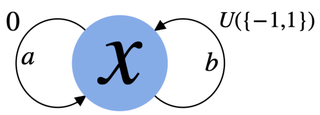
VECTOR INSTITUTE

UNIVERSITY OF TORONTO

Overview

Risk-sensitive reinforcement learning: two states can have the same value function, but very different return distributions.





How to address this? Risk measures: functions $\rho : \mathcal{P}(\mathbb{R}) \to \mathbb{R}$ which quantify the utility of different return distributions.

Goal of our work: Learn models which can be used to efficiently plan for risk-sensitive RL.

Existing methods:

1. Learn a model using MLE
   - ⇒ Does not take the decision problem into account.
   - ⇒ Inefficient when model does not have capacity to capture the *entire* environment.
2. Learn a model using (proper) value equivalence
   - ⇒ Learns a model by ensuring it matches the true value function over a set of policies: $\mathcal{M}^{\infty}(\Pi) = \left\{\tilde{m} : V^{\pi} = V^{\pi}_{\tilde{m}}, \quad \forall \pi \in \Pi\right\}$.
   - ⇒ Can perform arbitrarily poorly for risk measures other than expectation.

Our contributions:

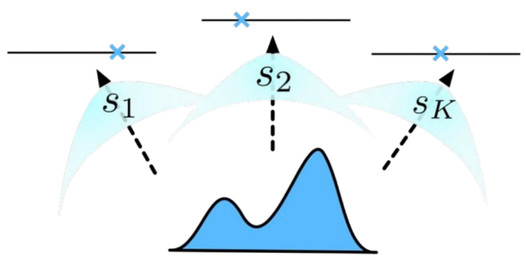
- We prove that proper value equivalence is not sufficient for risk-sensitive planning.
- We introduce distribution equivalence, a framework for learning models based on return distributions, and prove it is sufficient for planning for any risk measure, but is intractable in practice.
- We introduce statistical functional equivalence, a way of approximating distribution equivalence parameterized by a choice of functional.
- We demonstrate how our framework can be combined into any existing distributional model-free risk-sensitive RL algorithm.

Statistical functional equivalence

An ideal model for risk-sensitive reinforcement learning would match the true *return distribution* at each state. This would correspond to the set of models $\mathcal{M}^{\infty}_{\text{dist}}(\Pi) = \left\{\tilde{m} : \eta^{\pi} = \eta^{\pi}_{\tilde{m}}, \quad \forall \pi \in \Pi\right\}$.

This is too difficult to learn in practice; probability distributions are infinite-dimensional, and we must parameterize them in some finite-dimensional space for computational purposes.

This mapping is a statistical functional – it extracts certain aspects of a distribution as real values.



Statistical functional equivalence: the set of models which match the true *return statistics* (statistical functional of the return distribution) at each state.
Formally, the set of models $\left\{\tilde{m} : \boldsymbol{s}(\eta^{\pi}) = \boldsymbol{s}(\eta^{\pi}_{\tilde{m}}), \quad \forall \pi \in \Pi\right\}$.

Choice of statistical functional

The choice of the statistical functional has two important considerations: (i) how easily it can be learnt, and (ii) which risk measures it can plan for.

We provide theoretical guarantees on which statistical functionals can plan for which risk measures. For exact planning, it is sufficient that for any probability distribution ν, the risk measure $\rho(\nu)$ can be written as a linear combination of the statistical functional of ν: $\exists \alpha_0, \ldots, \alpha_m$ such that $\rho(\nu) = \alpha_0 + \sum_{i=1}^{m} \alpha_i \boldsymbol{s}_i(\nu)$.

We consider two sets of statistical functionals in particular for our experiments: the set of the first m moments, and the set of m evenly-spaced quantiles.

Combining with existing algorithms

A generic model-free distributional risk-sensitive RL algorithm uses a replay buffer $\mathcal{D}$ to learn a statistical functional $\boldsymbol{s}$ of the return for each state.
We can augment this algorithm with a statistical functional equivalent model $\tilde{m}$, by learning $\tilde{m}$ using the loss

$$\mathcal{L}_{\mathcal{D},\boldsymbol{s}}(\tilde{m}) = \mathop{\mathbb{E}}_{\substack{(x,a,r,x' \sim \mathcal{D}) \\ \tilde{x} \sim \tilde{m}(\cdot|x,a)}} \left[\sum_{i=1}^{m} \left(\boldsymbol{s}_i(x') - \boldsymbol{s}_i(\tilde{x}') \right)^2 \right].$$

Empirical results

We evaluate our framework across a collection of risky tabular environments and a continuous option trading domain.

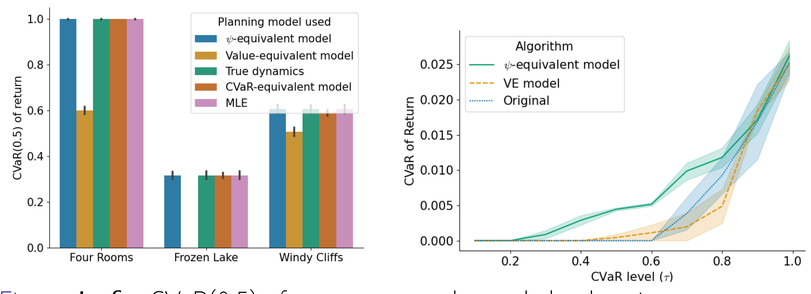


Figure: **Left**: CVaR(0.5) of returns across three tabular domains. **Right**: CVaR of returns for option trading domain.

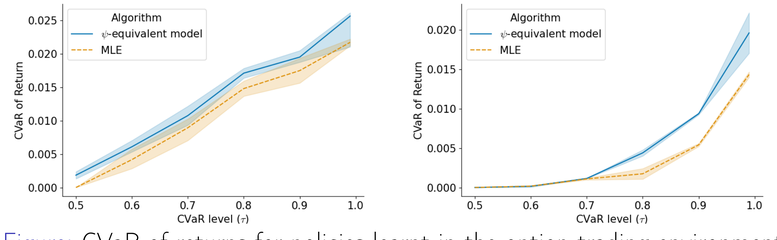


Figure: CVaR of returns for policies learnt in the option trading environment with distracting dimensions (**Left**: 2, **Right**: 6).

Our framework beats PVE for risk-sensitive planning, and beats MLE at modelling what is important for noisy environments.

CIFAR

NSERC CRSNG